IE6600 Computation and Visualization

Spring 2024 SEC 03

PROJECT 2

FINAL REPORT

Inmate Demographics Project

GROUP 3:

Lokhi Nalam (002649847)

Pooja Arumugam (002872003)

Sathvik Ramappa (002847460)

Arya Lokesh Gowda (002249418)

Part 1: Introduction

In the pursuit of advancing our skills in advanced data visualization and statistical data analysis, our project centered on the utilization of the Seaborn library. We aimed to explore and analyze distinct datasets from data.gov, delving into sectors such as health, environment, finance, or transportation. For the scope of this endeavor, we selected a dataset focused on inmates, obtained from the 'inmate.csv' file.

Our primary objective was to harness the capabilities of Seaborn to craft intricate static visualizations that not only unveil patterns and insights within the data but also serve as compelling tools for communication. Additionally, we sought to demonstrate our proficiency in saving these visualizations as image files, ensuring they are readily accessible for dissemination.

The culmination of our efforts involved the effective presentation of our analysis through a dynamic platform, be it a PowerPoint presentation or an engaging webpage. This multifaceted approach aimed to showcase the depth of our understanding and application of Seaborn for comprehensive data exploration and interpretation.

In the subsequent sections of this report, we will detail the various stages of our project, from data acquisition and inspection to the implementation of diverse Seaborn visualizations. Each visualization is a testament to our commitment to unraveling the intricacies of the inmate dataset, providing not only a visual narrative but also a profound understanding of the underlying statistical trends. Our journey through this project serves as a testament to our proficiency in leveraging Seaborn as a powerful tool for advanced data visualization and analysis.

Part 2: Dataset Selection and Confirmation

The 'inmate.csv' dataset was deliberately chosen due to its comprehensive nature, offering a wealth of information on inmate demographics, offenses, and legal aspects related to their incarceration. This dataset holds the promise of unveiling intricate patterns and trends within the criminal justice system, allowing for a deeper understanding of incarceration dynamics. Through meticulous data exploration, we aim to extract insights that highlight systemic disparities, illuminate law enforcement priorities, and elucidate societal challenges surrounding incarceration. Our commitment to meaningful analysis and interpretation drives us to create visualizations that not only showcase sophistication but also provide actionable insights for stakeholders. By leveraging this dataset, we aspire to contribute to informed discussions, advocate for evidence-based reforms, and foster a more equitable and effective criminal justice system. Ultimately, our goal is to harness the power of data to promote transparency, accountability, and positive societal change in the realm of incarceration and beyond.

Part 3: Data Acquisition and Inspection

The dataset exhibits diverse numerical characteristics, with notable variations in SID Number, TDCJ Number, Age, and Offense Code, capturing a range of inmate demographics and offense details. Categorical attributes like Gender, Race, Current Facility, County, and Parole Review Status add contextual dimensions to the inmate profiles. It is noteworthy that certain columns, such as Projected Release and Next Parole Review Date, contain missing values that may influence subsequent analyses. Overall, the dataset provides a comprehensive foundation for exploring inmate-related patterns and trends.

	SID Number	TDCJ Number	Name	Current Facility	Gender	Race	Age	Projected Release	Maximum Sentence Date	Parole Eligibility Date	Case Number	County	Offense Code	TDCJ Offense	Sentence Date	Offense Date	Sentence (Years)	Last Parole Decision	Next Parole Review Date	Parole Review Status
0	671628	2394062	ONOFRE,JESSE TINAJERO	Connally	M	H	88	04/23/2026	04/23/2026	04/22/2024	2019CR4680	Bexar	36010001	INDEC W/CHILD CONTACT	05/16/2022	08/02/2016	4.0	NaN	04/22/2024	IN PAROLE REVIEW PROCESS
1	686567	311644	PALACIOS,ROBERT LEONARD	Duncan	M	H	85	09/12/2028	05/12/2036	04/15/2022	320670	Harris	22100000	BURG HAB W/I SEXUAL ABUSE	10/21/1980	01/28/1980	30.0	Denied on 05/01/2023	04/2026	NOT IN REVIEW PROCESS
2	770626	449674	FLORES,ISABEL	Pack	M	H	86	01/01/9999	01/01/9999	09/02/2006	86CR-1234-B	Nueces	9150000	MURDER W/DEADLY WPN	04/13/1987	09/02/1986	Life	Denied on 07/21/2021	07/2024	NOT IN REVIEW PROCESS
3	771601	1491019	MOLETT,JOHN HENRY	W. Scott	M	B	81	12/02/2037	12/02/2037	12/02/2022	1144294	Harris	12990002	\GG ROBBERY W/DW	02/21/2008	12/03/2007	30.0	Denied on 12/28/2022	12/2023	IN PAROLE REVIEW PROCESS
4	799447	248098	DOWDEN,BILLY WAYNE	Hospital Galveston	M	W	86	01/01/9999	01/01/9999	05/24/1981	225031	Harris	9130000	CAPITAL MURDER	04/30/1975	06/28/1974	Capital Life	Denied on 11/20/2021	11/2024	NOT IN REVIEW PROCESS

Part 4: Data Cleaning and Preparation

Column Datatypes:

- We used the df.info() function to obtain information about the non-null records and data types of each column in the dataset. This allowed us to understand the nature of the data, identifying numerical, categorical, and date-related columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 132567 entries, 0 to 132566
Data columns (total 20 columns):
 #   Column                   Non-Null Count   Dtype 
---  ------                   --------------   ----- 
 0   SID Number               132567 non-null  int64 
 1   TDCJ Number              132567 non-null  int64 
 2   Name                     132567 non-null  object
 3   Current Facility         132567 non-null  object
 4   Gender                   132567 non-null  object
 5   Race                     132567 non-null  object
 6   Age                      132567 non-null  int64 
 7   Projected Release        132561 non-null  object
 8   Maximum Sentence Date    131898 non-null  object
 9   Parole Eligibility Date  120478 non-null  object
 10  Case Number              132488 non-null  object
 11  County                   132567 non-null  object
 12  Offense Code             132567 non-null  int64 
 13  TDCJ Offense             132567 non-null  object
 14  Sentence Date            129726 non-null  object
 15  Offense Date             132567 non-null  object
 16  Sentence (Years)         132482 non-null  object
 17  Last Parole Decision     69447 non-null   object
 18  Next Parole Review Date  105963 non-null  object
 19  Parole Review Status     120993 non-null  object
dtypes: int64(4), object(16)
memory usage: 20.2+ MB
```

- Date-related columns, such as 'Projected Release,' 'Maximum Sentence Date,' 'Parole Eligibility Date,' 'Sentence Date,' 'Offense Date,' and 'Last Parole Decision,' were converted to the datetime format to facilitate temporal analysis.

```
SID Number                          int64
TDCJ Number                         int64
Name                               object
Current Facility                 category
Gender                           category
Race                             category
Age                                 int64
Projected Release          datetime64[ns]
Maximum Sentence Date      datetime64[ns]
Parole Eligibility Date    datetime64[ns]
Case Number                        object
County                           category
Offense Code                        int64
TDCJ Offense                     category
Sentence Date              datetime64[ns]
Offense Date               datetime64[ns]
Sentence (Years)                  float64
Last Parole Decision       datetime64[ns]
Parole Review Status             category
dtype: object
```

Missing Values:

- The presence of missing values was assessed using the df.isnull().sum() function. We identified columns with null values and implemented appropriate strategies, such as dropping columns with more than 10% missing data and filling null records where necessary.

Duplicates:

- Duplicates in the dataset were identified and removed using the df.drop_duplicates() function, ensuring the integrity of our analysis.

Categorical Columns:

- Categorical columns, including 'Gender,' 'Race,' 'Current Facility,' 'County,' 'TDCJ Offense,' and 'Parole Review Status,' were explicitly converted to the categorical data type for more efficient memory usage and improved analysis.

Offense Code Refinement:

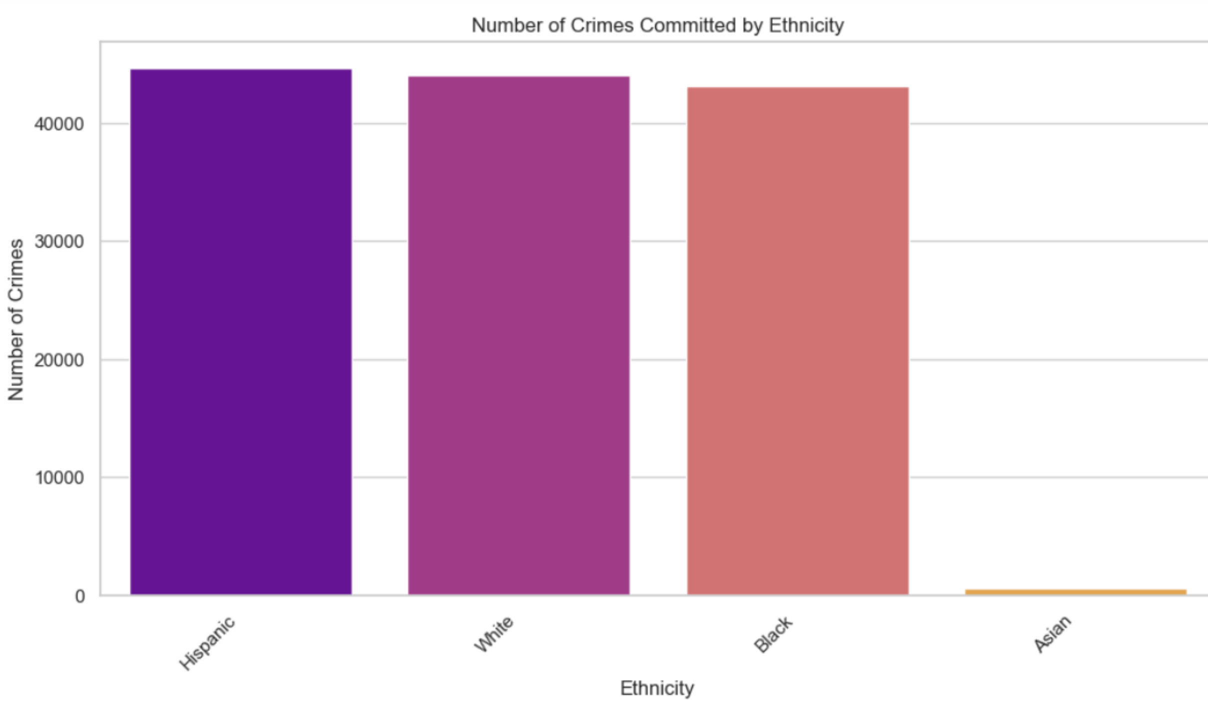
- To enhance the clarity of our analysis, we processed the 'Offense Code' column by cutting the last three decimal points, providing a more interpretable representation.
- We introduced a new column, 'Code Range,' by dividing the 'Offense Code' by 1,000,000. This division resulted in the creation of discrete code ranges that group offenses based on their numerical magnitude.
- Leveraging the 'TDCJ Offense' column, we mapped common words associated with each offense group. This step was crucial for providing a descriptive and human-readable label to each code range.
- We assigned serial numbers starting from 1 to the distinct offense groups. This sequential ordering aids in the interpretation and presentation of the data.
- The refined offense code groups were displayed, showcasing the mapping between code ranges and the corresponding offenses. Each group, identified by a serial number, represents a cluster of related offenses.

We discovered that the 'Offense Type' had a large number of unique values. To enhance understandability and simplify our analysis, we aimed to group these diverse offense types into broader categories or buckets. The modified 'Offense Code' column is used to create groups based on the range of offense codes. Common words from the 'TDCJ Offense' column are mapped to each group. The groups are assigned serial numbers (1, 2, 3, ...) and named based on the most common word in the 'TDCJ Offense' within each group. The final DataFrame, inmate_df, now includes a 'Code Range' column representing the grouped offenses and a new 'TDCJ Offense' column indicating the associated activity.

Part 5: Exploratory Data Analysis (EDA) Using Seaborn

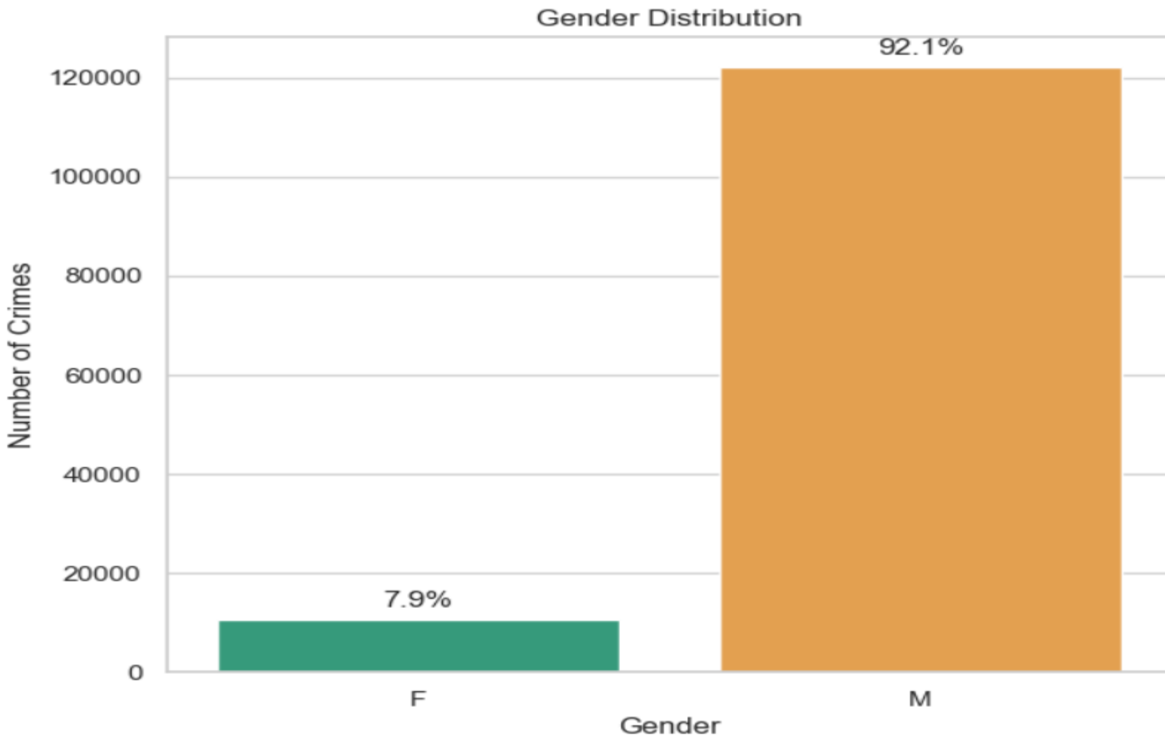
Crimes Committed by Ethnicity

The data reveals distinct patterns in criminal involvement across different ethnic groups. Hispanic individuals emerge with the highest number of reported crimes, totaling 44,661 instances, suggesting a significant proportion of offenders within the dataset. This finding may reflect various socio-economic factors, including disparities in access to resources, educational opportunities, and systemic biases within the criminal justice system. Following closely are White and Black individuals, with 44,003 and 43,118 reported crimes, respectively, indicating a comparable level of involvement in criminal activities. The relatively lower numbers of crimes attributed to Asian, Other, and Indigenous individuals may reflect smaller population sizes within the dataset or potentially different socio-cultural dynamics. Notably, the small number of instances where ethnicity is unknown underscores the importance of data completeness and accuracy for robust analysis. These findings highlight the complex interplay of socio-economic, cultural, and systemic factors influencing crime rates among diverse ethnic groups, warranting further exploration and targeted interventions to address underlying disparities and promote social equity.



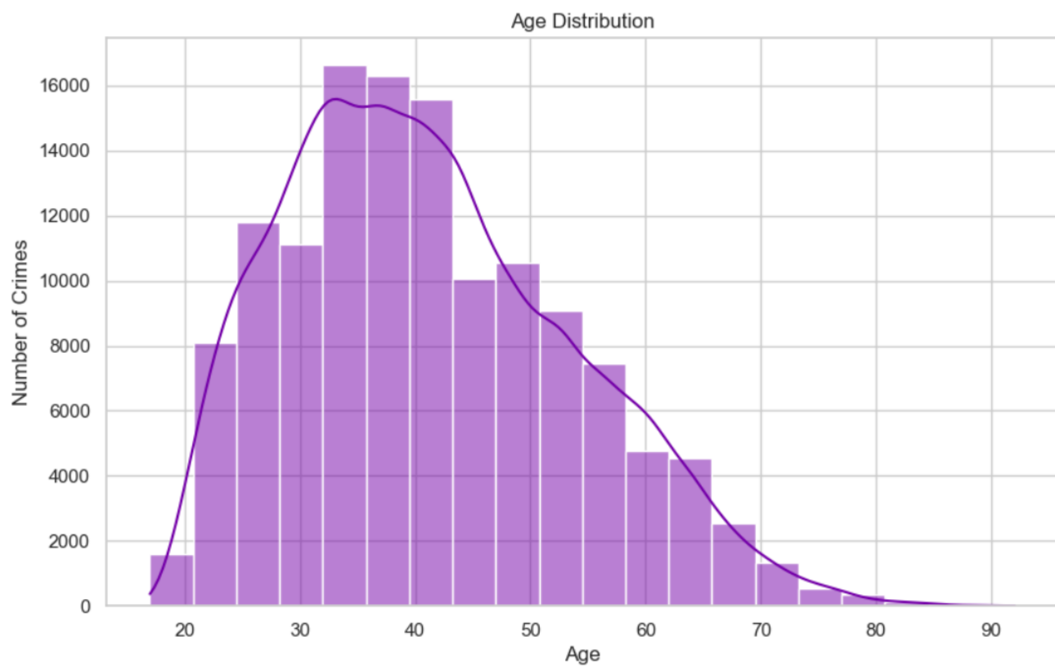
Gender Distribution

- Analysis of gender distribution reveals a predominant representation of males, with 122,045 instances, compared to females, which account for 10,522 instances within our dataset.
- This insight can be valuable for understanding the demographic composition of individuals involved in the reported offenses.



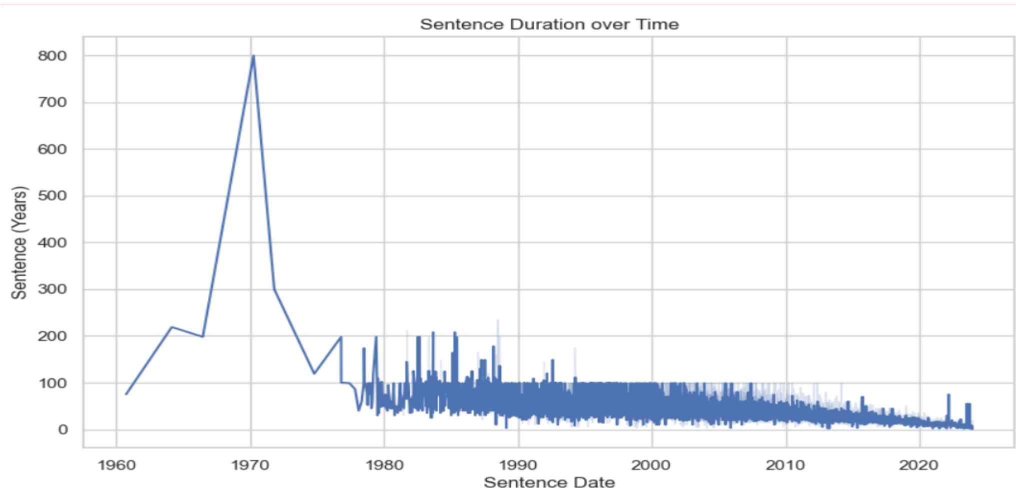
Analyzing Age Distribution

- There is a higher concentration of individuals in their 30s, as indicated by the peak in the histogram.
- The distribution generally shows a gradual decline as age increases, with fewer individuals in older age groups.
- The line overlaid on the histogram is the Kernel Density Estimate (KDE). It provides a smoothed representation of the distribution, offering insights into the overall shape and trends in age distribution.



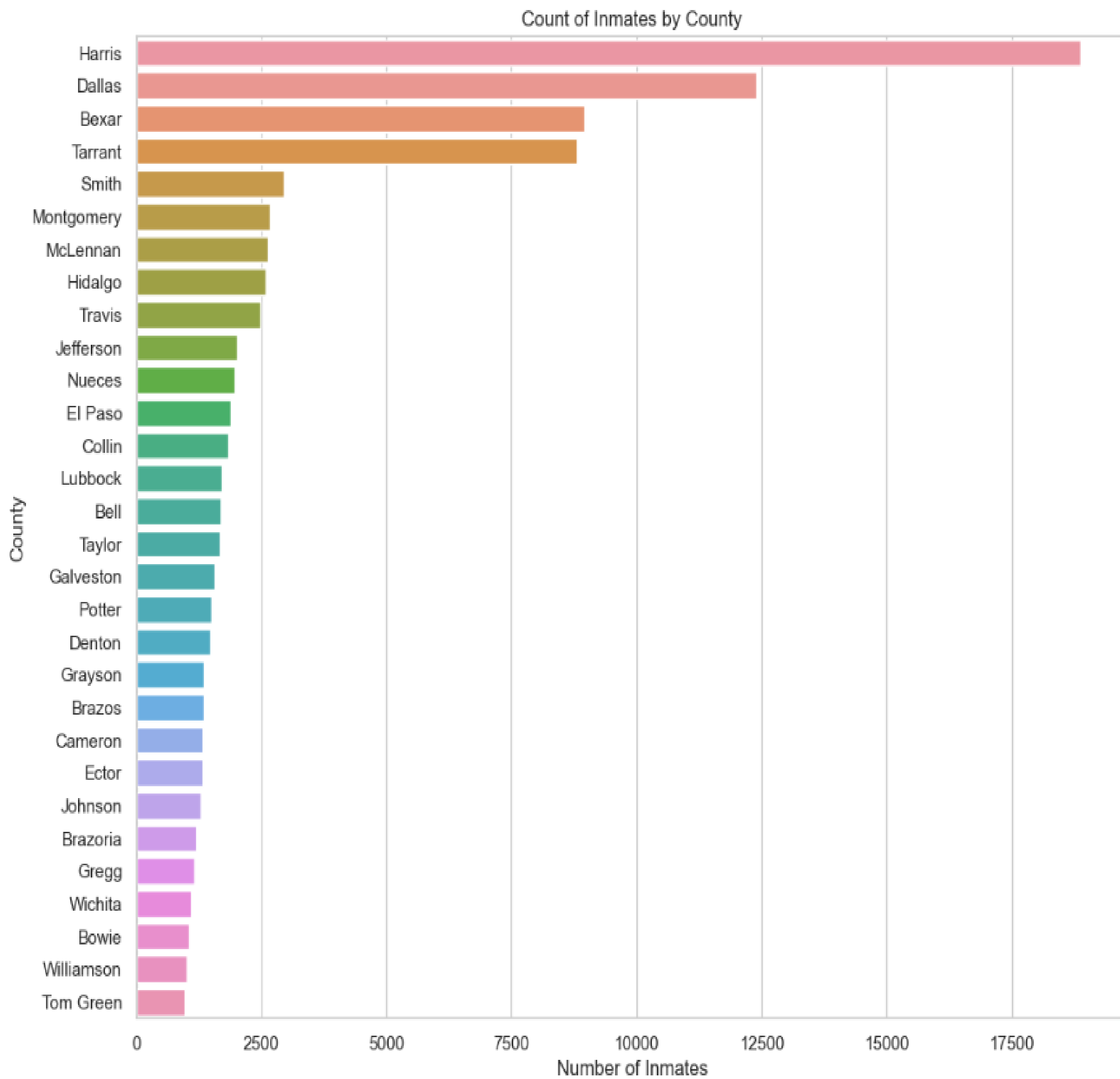
Sentence Duration over Time

- The plot reveals the diversity in sentence durations across different time periods.
- Identifying trends or patterns in sentence durations over time may offer insights into legal and judicial practices evolution.
- The visualization allows a quick scan for missing sentence duration values, guiding data quality assessment.



Inmate Distribution Across Top Counties

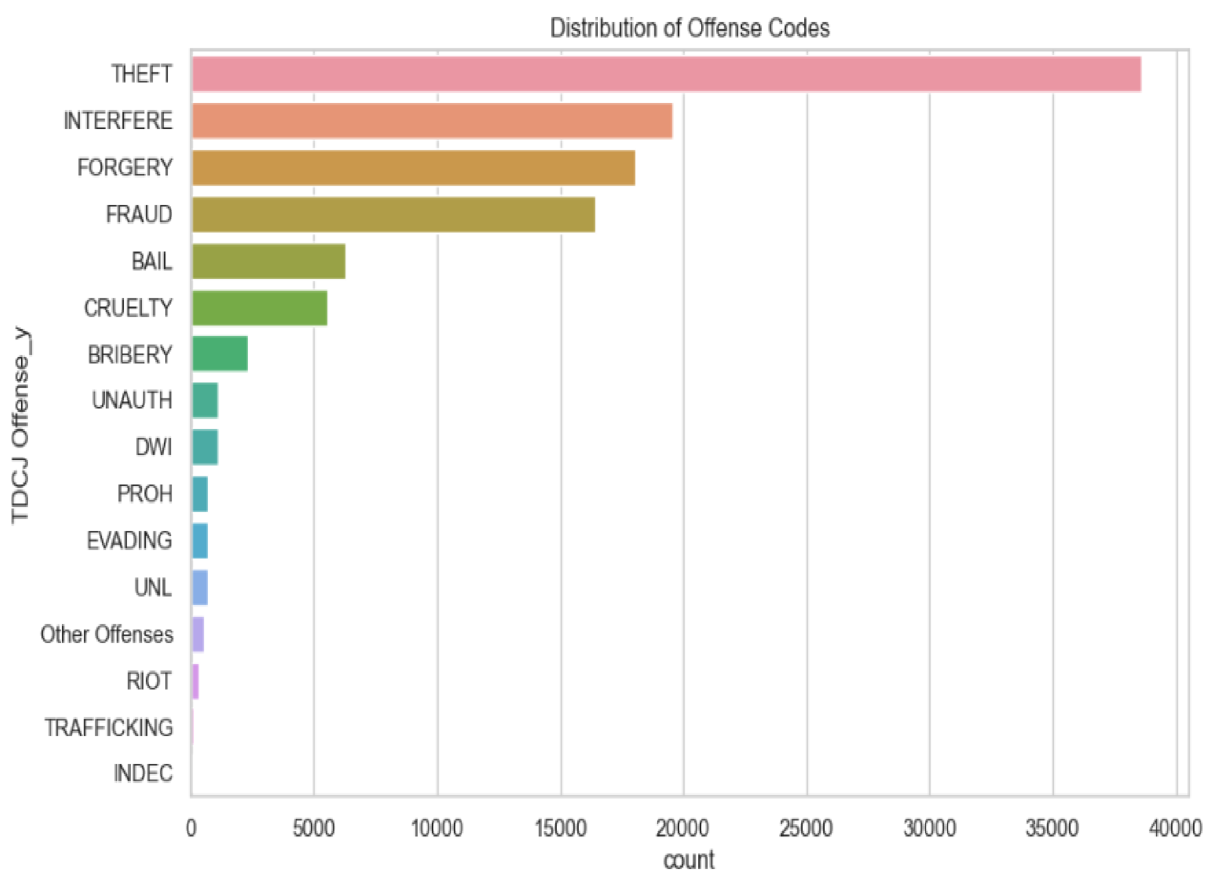
- Harris County leads with 18,877 inmates, signifying its significant role in the dataset.
- Other major contributors include Dallas (12,399), Bexar (8,963), Tarrant (8,813), and Smith (2,960), demonstrating substantial inmate populations.
- Counties with comparatively lower inmate populations, include Tom Green (978), Williamson (1,007), and Bowie (1,056).



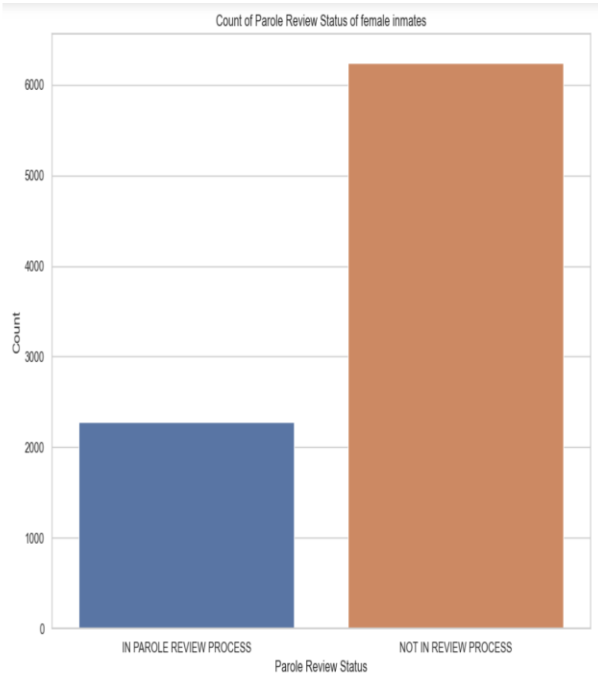
Offense Codes Distribution

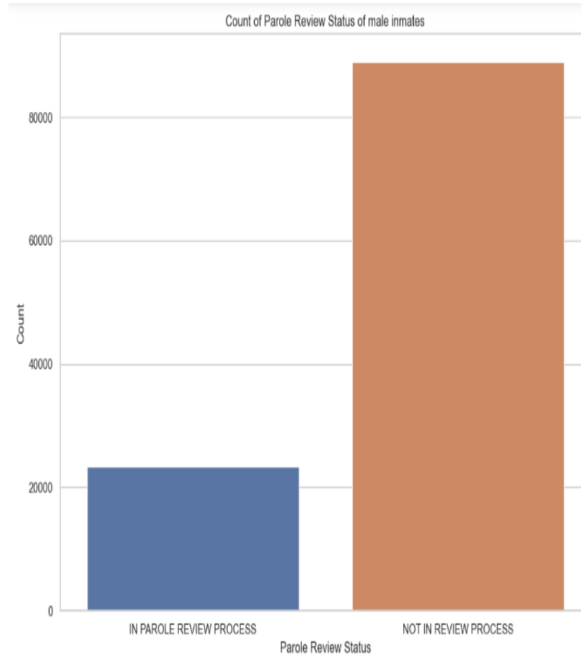
- The data provides a breakdown of TDCJ (Texas Department of Criminal Justice) offenses and their respective frequencies within the dataset. Here's a brief analysis of the findings:
- Theft (38,591 instances): Theft appears to be the most prevalent offense, indicating a significant occurrence within the dataset. This may suggest various factors such as economic conditions, opportunity, and criminal motivations prevalent in the studied population.
- Interference (19,548 instances): Interference follows closely, indicating a substantial number of cases involving obstruction or interference with legal processes, law enforcement, or public administration.
- Forgery (18,061 instances): Forgery is another prevalent offense, indicating instances of fraudulently altering or creating documents for deceptive purposes.
- Fraud (16,450 instances): Fraud denotes a considerable number of cases involving deceit, misrepresentation, or deception for financial gain, suggesting a significant concern within the dataset.
- Bail (6,281 instances): Bail-related offenses involve violations or issues related to bail conditions or bail bonds, indicating a notable presence in the dataset.
- Cruelty (5,551 instances): Cruelty offenses may involve acts of violence, abuse, or mistreatment, indicating a concerning occurrence within the dataset.
- Bribery (2,301 instances): Bribery offenses suggest instances of offering or receiving bribes for personal gain, highlighting potential corruption or unethical behavior.
- Unlawful (1,139 instances): Unlawful acts may encompass a range of illegal activities not specified in other categories, indicating a diverse spectrum of offenses.
- DWI (Driving While Intoxicated) (1,110 instances): DWI offenses involve operating a motor vehicle while under the influence of alcohol or drugs, indicating a notable occurrence within the dataset.
- Prohibited (706 instances): Prohibited acts may refer to various activities prohibited by law, regulations, or legal orders, suggesting a range of offenses falling under this category.
- Evading (693 instances): Evading offenses may involve attempts to evade law enforcement or legal obligations, indicating potential resistance or avoidance behavior.

- Unlawful (678 instances): Unlawful acts may encompass a range of illegal activities not specified in other categories, indicating a diverse spectrum of offenses.
- Other Offenses (560 instances): Other offenses encompass a broad category of offenses not explicitly listed, suggesting a need for further categorization or analysis to understand their nature and prevalence.
- Riot (339 instances): Riot offenses involve acts of violence, disorderly conduct, or public disturbances involving multiple individuals, indicating potential social or civil unrest within the dataset.
- Trafficking (136 instances): Trafficking offenses involve illegal trade or transportation of goods, persons, or substances, indicating instances of organized criminal activity.
- Indecency (69 instances): Indecency offenses may involve acts of impropriety, lewd behavior, or sexual misconduct, indicating potential violations of societal norms or legal standards.



Parole Review Status Across Genders





Female Distribution:

Around 5,000 of the female population are in parole review process and around 15000 of the population are not in review process.

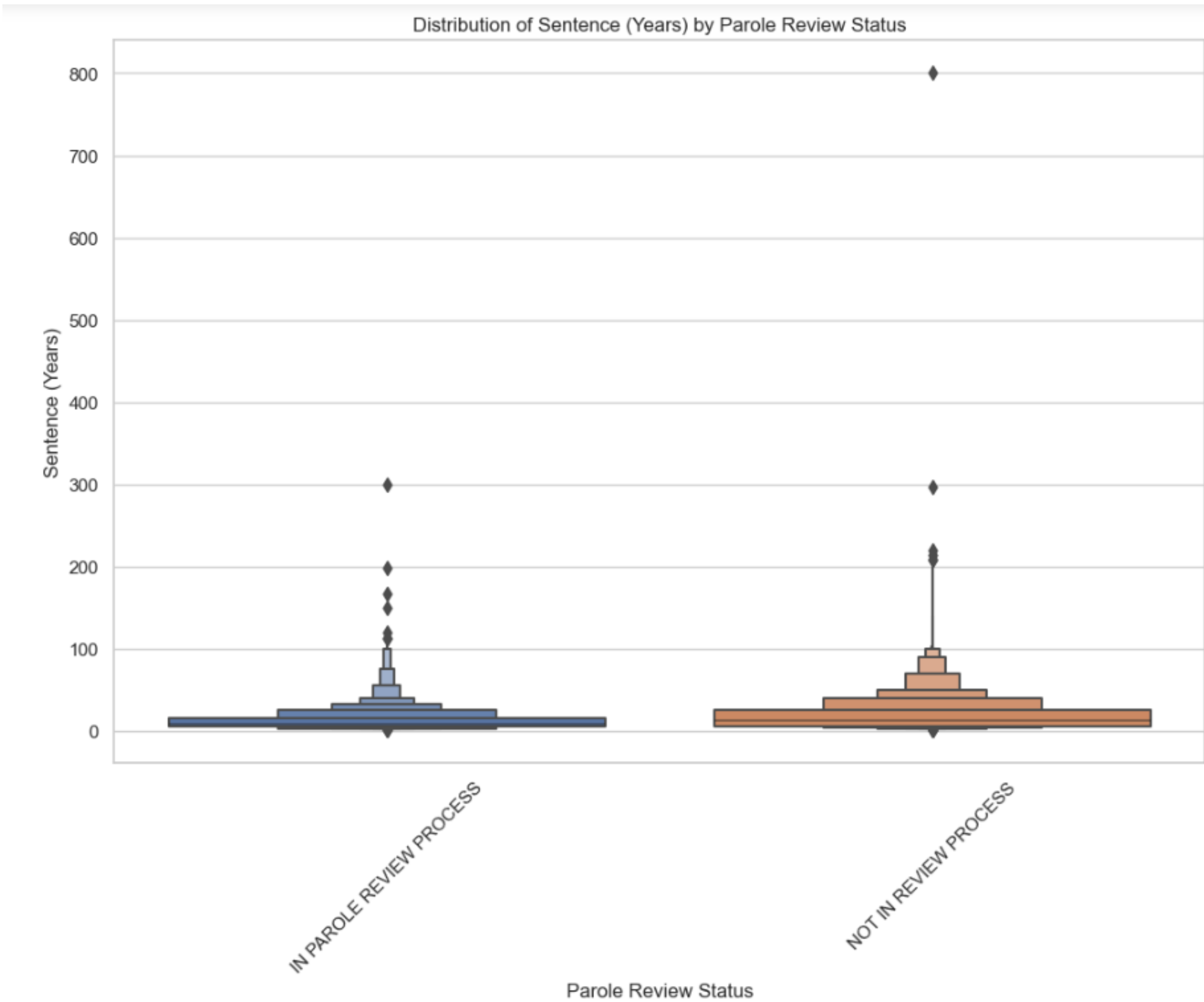
Male Distribution:

Around 21,000 of the male population are in parole review process and around 87000 of the population are not in review process.

Approximately 30,522 individuals, across both genders, are actively undergoing the parole review process.

Sentence by Parole Review Status Distribution

- Inmates not in the parole review process tend to have longer sentences, with outliers exceeding 800 years.
- Those in the parole review process show a more varied distribution, with some outliers indicating sentences up to 300 years.
- Understanding sentence duration disparities across parole review statuses is essential for informed decision-making in the correctional system.



Count plot of inmates by Current Facility

- Coffield, Allred, and Beto facilities emerge as the top three facilities with the highest inmate populations, indicating their significance within the Texas Department of Criminal Justice (TDCJ) system. These facilities likely handle a substantial portion of the incarcerated population and may require focused resources and management attention.
- Estelle and Robertson facilities closely follow in terms of inmate population, suggesting they are also significant facilities within the TDCJ system. Understanding the characteristics and operations of these facilities can provide insights into the broader dynamics of inmate housing and management.
- The distribution highlights a considerable variation in the sizes of TDCJ facilities, with larger facilities such as Coffield and Allred accommodating significantly more inmates compared to smaller facilities like West Texas Hospital and Santa Maria Baby Bonding. This variation may reflect differences in facility capacities, security levels, and specialized services provided.
- Facilities with higher inmate populations may require more resources in terms of staffing, infrastructure, and programming to effectively manage and meet the needs of the incarcerated population. Understanding the distribution of inmates across facilities can inform resource allocation decisions within the TDCJ system.
- The presence of facilities such as Hospital Galveston, Goodman, Baten, and West Texas Hospital suggests the existence of specialized facilities catering to specific needs such as healthcare, mental health services, or specialized populations. These facilities play a crucial role in addressing the diverse needs of the inmate population and ensuring access to appropriate care and services.
- Analyzing facility distribution can inform policy discussions around inmate housing, facility capacity planning, and the allocation of resources within the TDCJ system. Policy decisions aimed at improving efficiency, effectiveness, and outcomes within the correctional system can benefit from a comprehensive understanding of facility utilization and inmate population distribution.

Overall, the distribution of inmates across TDCJ facilities provides valuable insights into the scale, diversity, and dynamics of the correctional system, guiding decision-making processes and resource allocation efforts aimed at enhancing operational effectiveness and meeting the needs of incarcerated individuals.